Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study

Staartjes, Victor E ; Serra, Carlo ; Muscas, Giovanni ; Maldaner, Nicolai ; Akeret, Kevin ; van Niftrik, Christiaan H B ; Fierstra, Jorn ; Holzmann, David ; Regli, Luca

Abstract: OBJECTIVEGross-total resection (GTR) is often the primary surgical goal in transsphenoidal surgery for pituitary adenoma. Existing classifications are effective at predicting GTR but are often hampered by limited discriminatory ability in moderate cases and by poor interrater agreement. Deep learning, a subset of machine learning, has recently established itself as highly effective in forecasting medical outcomes. In this pilot study, the authors aimed to evaluate the utility of using deep learning to predict GTR after transsphenoidal surgery for pituitary adenoma.</jats:sec><jats:sec>METHODSData from a prospective registry were used. The authors trained a deep neural network to predict GTR from 16 preoperatively available radiological and procedural variables. Class imbalance adjustment, cross-validation, and random dropout were applied to prevent overfitting and ensure robustness of the predictive model. The authors subsequently compared the deep learning model to a conventional logistic regression model and to the Knosp classification as a gold standard.</jats:sec><jats:sec>RESULTSOverall, 140 patients who underwent endoscopic transsphenoidal surgery were included. GTR was achieved in 95 patients (68%), with a mean extent of resection of 96.8% ± 10.6%. Intraoperative high-field MRI was used in 116 (83%) procedures. The deep learning model achieved excellent area under the curve (AUC; 0.96), accuracy (91%), sensitivity (94%), and specificity (89%). This represents an improvement in comparison with the Knosp classification (AUC: 0.87, accuracy: 81%, sensitivity: 92%, specificity: 70%) and a statistically significant improvement in comparison with logistic regression (AUC: 0.86, accuracy: 82%, sensitivity: 81%, specificity: 83%) (all p lt; 0.001).</jats:sec><jats:sec>CONCLUSIONSIn this pilot study, the authors demonstrated the utility of applying deep learning to preoperatively predict the likelihood of GTR with excellent performance. Further training and validation in a prospective multicentric cohort will enable the development of an easy-to-use interface for use in clinical practice.

DOI: https://doi.org/10.3171/2018.8.focus18243

NEUROSURGICAL FOCUS

Neurosurg Focus 45 (5):E12, 2018

Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study

***Victor E. Staartjes, BMed,[1] Carlo Serra, MD,[1] Giovanni Muscas, MD,[2] Nicolai Maldaner, MD,[1] Kevin Akeret, MD,[1] Christiaan H. B. van Niftrik, MD,[1] Jorn Fierstra, MD, PhD,[1] David Holzmann, MD,[3] and Luca Regli, MD[1]**

[1]Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Switzerland; [2]Department of Neurosurgery, Tuscany School of Neurosurgery, University of Firenze, Italy; and [3]Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich, University of Zurich, Switzerland

OBJECTIVE Gross-total resection (GTR) is often the primary surgical goal in transsphenoidal surgery for pituitary adenoma. Existing classifications are effective at predicting GTR but are often hampered by limited discriminatory ability in moderate cases and by poor interrater agreement. Deep learning, a subset of machine learning, has recently established itself as highly effective in forecasting medical outcomes. In this pilot study, the authors aimed to evaluate the utility of using deep learning to predict GTR after transsphenoidal surgery for pituitary adenoma.

METHODS Data from a prospective registry were used. The authors trained a deep neural network to predict GTR from 16 preoperatively available radiological and procedural variables. Class imbalance adjustment, cross-validation, and random dropout were applied to prevent overfitting and ensure robustness of the predictive model. The authors subsequently compared the deep learning model to a conventional logistic regression model and to the Knosp classification as a gold standard.

RESULTS Overall, 140 patients who underwent endoscopic transsphenoidal surgery were included. GTR was achieved in 95 patients (68%), with a mean extent of resection of 96.8% ± 10.6%. Intraoperative high-field MRI was used in 116 (83%) procedures. The deep learning model achieved excellent area under the curve (AUC; 0.96), accuracy (91%), sensitivity (94%), and specificity (89%). This represents an improvement in comparison with the Knosp classification (AUC: 0.87, accuracy: 81%, sensitivity: 92%, specificity: 70%) and a statistically significant improvement in comparison with logistic regression (AUC: 0.86, accuracy: 82%, sensitivity: 81%, specificity: 83%) (all $p < 0.001$).

CONCLUSIONS In this pilot study, the authors demonstrated the utility of applying deep learning to preoperatively predict the likelihood of GTR with excellent performance. Further training and validation in a prospective multicentric cohort will enable the development of an easy-to-use interface for use in clinical practice.

https://thejns.org/doi/abs/10.3171/2018.8.FOCUS18243

KEYWORDS pituitary surgery; deep learning; deep neural network; outcome prediction; pituitary adenoma; transsphenoidal surgery

Over the past few decades, the transnasal transsphenoidal approach has become the preferred technique for the resection of most pituitary adenomas (PAs).[10] Using either endoscopic or microscopic techniques, excellent surgical and endocrinological results can be achieved with minimal morbidity and mortality.[5,6] In the majority of cases, gross-total resection (GTR) can be achieved.[7] Especially in patients harboring secreting adenomas, where biochemical cure is targeted, GTR is the surgical goal. Subtotal resection and revision surgery have been linked to excess morbidity and mortality.[5,16,17,18] Through the development of assistive techniques such as intraoperative high-field MRI, the rates of GTR have been steadily increasing.[20,24]

ABBREVIATIONS AUC = area under the curve; CSS = cavernous sinus space; EOR = extent of resection; GTR = gross-total resection; ICD = intercarotid distance; NPV = negative predictive value; PA = pituitary adenoma; PPV = positive predictive value; 3T-iMRI = 3-T intraoperative MRI.

SUBMITTED May 19, 2018. **ACCEPTED** August 20, 2018.
INCLUDE WHEN CITING DOI: 10.3171/2018.8.FOCUS18243.

* V.E.S. and C.S. contributed equally to this work.

Predictive analytics for GTR may help in surgical decision-making. Patients with severe comorbidities may profit more from nonsurgical forms of treatment if GTR is highly unlikely. This is particularly true for recurrent functioning adenomas. Lastly, an accurate predictive model is helpful in preoperative patient counseling.[1]

The likelihood of achieving GTR is influenced by a wealth of factors, including invasion into the cavernous sinus space (CSS) and dura, PA diameters and volumes, and growth patterns, as well as sellar anatomy.[7,8,13,14] Since integrating the complex interactions among all these predictive factors for subtotal resection in daily clinical practice is not feasible, morphological classifications such as those developed by Knosp[11] and Hardy[9] have been introduced. These classifications are valuable for predicting the likelihood of GTR. The highest accuracy is observed in extreme cases of overt invasion or noninvasion, such as encasement of the internal carotid artery or in very small intrasellar PAs.[7,14] However, the sensitivity and specificity of these grading systems are low in intermediate cases, for which surgeons are most interested in having an analytical model for predicting GTR.[7,14] Furthermore, they have demonstrated low interrater agreement.[15]

Currently, machine learning is being implemented in clinical practice at an increasing rate to improve predictive power over conventional statistical methods. Indeed, several machine learning methods have demonstrated high predictive ability for neurosurgical data.[2,19] Deep learning represents a further development of machine learning. A subgroup of artificial neural networks, which has not yet been applied extensively in neurosurgery, has proven itself as an excellent analytical method in other disciplines.[2,12] Instead of relying on a single level of abstraction, as many machine learning models do, deep neural networks have a structure that decomposes complex problems into multiple simpler ones.[12] This enables discovering intricate relationships between variables while also automatically selecting only the most important input variables, often resulting in improved predictive ability.[12] Our aim was to evaluate the feasibility and utility of predicting GTR in transsphenoidal surgery for PA using deep learning in a pilot study.

Methods

Patient Population

A consecutive series of patients who underwent endoscopic transnasal transsphenoidal surgery for PA performed by 2 senior neurosurgeons (L.R. and C.S.) at the Department of Neurosurgery, University Hospital Zurich, was evaluated. To be included, patients had to have complete preoperative as well as 3-month postoperative neuroimaging data. Exclusion criteria were transcranial or combined procedures, as well as those planned for a limited decompression only. From October 2012 onward, all patients were treated according to the same PA protocol as previously described.[20] The preoperative surgical goal for each adenoma was set based on the invasiveness pattern. Adenomas classified as Knosp grade 0, 1, or 2 were initially considered noninvasive.[7,14] Whenever safely possible, GTR was attempted even in cases deemed invasive. Clinical and radiological data were collected in a prospective registry. High-field 3-T intraoperative MRI (3T-iMRI) was routinely performed unless contraindicated. Data were treated according to the ethical standards of the Declaration of Helsinki. The registry was approved by our institutional committee.

Outcome Measures

We defined GTR at 3 months as our primary endpoint. Patients underwent preoperative and 3-month postoperative volumetric contrast-enhanced MRI (3-T Skyra VD13, Siemens) at a field strength of 3 T. Rating was performed by a board-certified neurosurgeon with extensive experience in pituitary surgery and imaging. Adenoma morphology was graded according to the modified Knosp[14] and Hardy[9] classifications. Each adenoma was also manually contoured on source volumetric sequences to allow subsequent 3D rendering and volumetric measurement through the software (iPlan Cranial, Brainlab). Extent of resection (EOR) was measured on 3-month postoperative MRI and was calculated as the percentagewise reduction of residual tumor volume to baseline tumor volume on preoperative MRI. An EOR of 100% corresponded to GTR. The smallest distance between the 2 horizontal C_4 segments of the internal carotid arteries was defined as the intercarotid distance (ICD), and tumor diameters in 3 axes were obtained on coronal sections.[3,21]

Statistical Analysis

Table 1 provides an explanation of the most important concepts in machine learning–based outcome prediction. Extended methods and model specifications can be found in Appendix 1. Missing data were completed using predictive mean matching. To counteract class imbalance, the synthetic minority oversampling technique was applied.[4] We considered the modified Knosp classification as the gold standard for predicting GTR and applied the commonly used threshold for CSS invasion (grades 3A and higher) for binomial classification.[7,14] Thus, adenomas with Knosp grade 0, 1, or 2 were preoperatively deemed to be completely resectable. Confusion matrices were generated to obtain accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score in predicting GTR. The area under the curve (AUC) was calculated using the nonbinomial modified Knosp classification with 6 grades.[14]

We then compared the modified Knosp classification's performance against deep learning and against logistic regression. To extract the optimal training performance from our relatively small data set, we used 5-fold cross-validation without holdout to assess out-of-sample performance for deep learning and logistic regression. Hyperparameters were tuned to find the most robust models in terms of AUC. The best model was then selected, and final performance measures along with 95% confidence intervals were obtained by repeated cross-validation. The following variables were included in the models: sex; age; prior transsphenoidal surgery; Knosp and Hardy classifications;[9,14] invasiveness;[7,14] intercarotid distances at the C_6, C_4 horizontal, and C_4 vertical segments as defined by Bouthillier et al.;[3] the R ratio between maximum adenoma

TABLE 1. Definitions of the most important concepts in machine learning–based outcome prediction

Concept	Definition
ML	Computer-based methods for classification (prediction of classes) and regression (prediction of values) that improve by minimizing a prespecified error function. Common ML methods are neural networks, gradient boosting, decision trees, Naive Bayes, and support vector machines. ML methods can further be categorized as supervised (models are trained based on known labels) or unsupervised (models discern patterns in the absence of known labels).
DL	A subset of machine learning based on neural networks (often MLPs) with more than 3 layers (deep). Similar to real neurons, these models learn by adjusting the reactivity of their neurons to certain inputs. DL can be applied as a supervised or unsupervised technique.
Hyperparameter tuning	Hyperparameters, which specify how a model learns, need to be set by the data scientist before training. They are perpetually improved (tuned) to find the model that performs best.
Imputation & multiple imputation	Methods to impute missing data in studies. Imputation enables retaining statistical power even when small to moderate amounts of data are missing. Multiple imputation is the current state of the art.
Class imbalance	When training data consist mainly of 1 class, the majority class, ML models perform poorly. This is because they can often achieve high accuracy easily by always predicting the majority class and not learning how to even recognize observations that would belong to the minority class. Neurosurgical data are often prone to class imbalance, e.g., when predicting complications, which occur in only 10%. Here, a model could simply always predict "no complication" (accuracy: 90%, specificity: 100%). This results in synthetically high accuracy, specificity, and AUC, but unemployable sensitivity. This is coined the "accuracy paradox."
SMOTE	The adverse effects of class imbalance can be negated by oversampling (collecting more observations of the minority class) or undersampling (cutting observations from the majority class). A state-of-the-art technique for oversampling is SMOTE, which synthesizes new observations for the minority class by averaging multiple other observations using an ML method called k-nearest neighbors. The resulting data set with reduced class imbalance can then be used for training, forcing the model to learn how to distinguish classes.
OSE	It is crucial to test the performance of a model on new data, previously unseen by the model during training, as this allows spotting overfitting to the training data. This performance is termed OSE.
k-fold CV	Data are randomly split into k folds, of which $k - 1$ folds are used for training and the model is evaluated on the remaining fold. This process is repeated k times; thus, k similar models are trained and evaluated. The mean performance metrics of the k models are then obtained. This method enables the use of all data for training, while also allowing an unbiased assessment of model performance; $k = 5$ is commonly used.
Holdout	Sometimes, when enough data are available, some are "sacrificed" (not used for training or validation) to provide an additional, unbiased assessment of model performance (holdout/testing set).
Overfitting	A phenomenon where a model overtrains and starts memorizing the observations used for training. This drastically reduces generalizability—the model will be unable to make accurate predictions on new observations. Overfitting is diagnosed by excellent performance on the training set and poor performance during (cross) validation (high OSE).
Dropout	A method in deep learning that randomly drops inputs to neurons, so that the model can never rely on any particular combination of inputs. This drastically reduces overfitting.
Confusion matrix	A table comparing the labels predicted by a classification model to the true labels. From this table, performance metrics can be calculated.
AUC	Integral-based performance metric. An area of 1.0 represents a perfect test; an area of 0.5 represents a worthless test. It enables assessment of predictive ability, and identification of an optimal threshold to distinguish between classes.
Accuracy	Proportion of true predictions (positive and negative) among all predictions.
Sensitivity	Proportion of correctly predicted positives among all true positives.
Specificity	Proportion of correctly predicted negatives among all true negatives.
PPV	Proportion of correctly predicted positives among all positive predictions; also termed "precision."
NPV	Proportion of correctly predicted negatives among all negative predictions.
F1 score	Composite metric defined as the harmonic mean of PPV and sensitivity.

CV = cross-validation; DL = deep learning; ML = machine learning; MLP = multilayer perceptron; OSE = out-of-sample error; SMOTE = synthetic minority oversampling.

diameter and ICD C_4 horizontal segment; availability of 3T-iMRI; and adenoma secretory status, volume, and diameters in 3 axes.

For deep learning, a multilayer perceptron with 5 hidden layers was trained in Keras (https://keras.io) using a TensorFlow (Google Brain Team, Google LLC) back end. Random dropout layers were implemented to minimize overfitting, and all predictors were included.[22] For comparison with conventional statistical methods, a standard logistic regression model including all predictors was trained and evaluated using cross-validation. We statistically compared deep learning and logistic regression performance using Welch's 2-sample t-test. All analyses were carried out in R version 3.4.4 (The R Foundation for Statistical Computing). Two-tailed tests were considered significant at $p \leq 0.05$.

TABLE 2. Baseline patient characteristics of the 140 included patients

Characteristic	Value
Female sex, n (%)	61 (44)
Mean age, yrs	54.0 ± 17.1
Prior surgery, n (%)	14 (10)
Tumor type, n (%)	
Nonfunctioning	95 (68)
GH-secreting	29 (21)
Prolactin-secreting	11 (8)
ACTH-secreting	3 (2)
TSH-secreting	1 (1)
Plurihormonal	1 (1)
Mean tumor diameter, mm	
X axis	21.5 ± 9.0
Y axis	17.1 ± 7.3
Z axis	21.0 ± 10.5
Mean ICD, mm	
ICD C_4 horizontal segment	21.1 ± 2.6
ICD C_4 vertical segment	17.1 ± 3.0
ICD C_6	14.1 ± 2.8
Mean R ratio	1.0 ± 0.4
Mean baseline tumor vol, cm^3	5.9 ± 7.6

ACTH = adrenocorticotropic hormone; GH = growth hormone; TSH = thyroid-stimulating hormone.
Mean values are presented as the mean ± SD.

Results

Patient Population

A total of 140 patients were included. Baseline patient characteristics are provided in Table 2; 3T-iMRI was used in 116 (83%) procedures. Overall, GTR was achieved in 95 patients (68%), with a mean EOR (± SD) of 96.8% ± 10.6% (Table 3). The mean residual tumor volume in the overall cohort was 0.31 ± 1.58 cm^3. Before multiple imputation, 89% of non-endpoint data fields were complete.

Knosp Classification

The Knosp classification (Table 4) scored well in terms of AUC (0.87), accuracy (81%), and F1 score (83%) and provided a sensitivity of 92% and an NPV of 90%. However, specificity (70%) and PPV (76%) were moderate (Table 5).

TABLE 3. Surgical results at the 3-month postoperative follow-up

Characteristic	Value
Use of 3T-iMRI, n (%)	116 (83)
GTR, n (%)	95 (68)
EOR (%)	
Median (IQR)	100 (98.6–100)
Mean ± SD	96.8 ± 10.6
Residual tumor vol, cm^3	
Median (IQR)	0.0 (0.0–0.1)
Mean ± SD	0.31 ± 1.58
Residual tumor in CSS, n (%)	31 (22)

TABLE 4. Adenoma morphology according to the modified Knosp and Hardy classifications

Morphology	Frequency
Knosp grade, n (%)	
0	25 (18)
1	29 (21)
2	45 (32)
3A	24 (17)
3B	9 (6)
4	8 (6)
Hardy grade (sellar), n (%)	
0	2 (1)
I	16 (11)
II	41 (29)
III	18 (13)
IV	63 (45)

Logistic Regression

Training resulted in an effective logistic regression model. On average, AUC values of 0.86, high accuracy, and an F1 score of 82% put the logistic regression model in range with the Knosp classification. The logistic regression model provided superior specificity (83%) and PPV (83%), but had a lower sensitivity (81%) and NPV (81%) than the Knosp classification. Knosp grade, invasiveness, 3T-iMRI, secretory status, and prior surgery were the significant predictors in the logistic regression models (all $p < 0.05$).

Deep Neural Network

After extensive hyperparameter optimization and elimination of overfitting using dropout, a powerful and robust deep learning model was obtained.[22] Figure 1 gives an illustrative overview of important features. Compared to the Knosp classification, AUC (0.96) and accuracy (91%) demonstrated a 10% increase. Specificity (89%) increased by nearly 20%, with comparable sensitivity of 94%. We also observed a marked increase in PPV (89%), NPV (94%), and F1 score (91%).

Due to repeated cross-validation, we were able to statistically test differences in performance metrics among deep learning and logistic regression. The deep learning model performed significantly better than logistic regression in terms of AUC (intergroup difference [Δ] 0.101, 95% CI 0.086–0.117), accuracy (Δ 0.089, 95% CI 0.077–0.099), sensitivity (Δ 0.128, 95% CI 0.111–0.145), specificity (Δ 0.056, 95% CI 0.035–0.077), PPV (Δ 0.055, 95% CI 0.035–0.075), NPV (Δ 0.125, 95% CI 0.092–0.158), and F1 score (Δ 0.091, 95% CI 0.079–0.104) (all $p < 0.001$).

Figure 2 shows the accuracy of the 3 prediction methods stratified by the Knosp classification.

TABLE 5. Performance metrics of the 3 predictive methods for GTR

Metric	Knosp Classification	Deep Neural Network		Logistic Regression		p Value
		Mean	95% CI	Mean	95% CI	
AUC	0.868	0.962	0.960–0.963	0.860	0.845–0.875	<0.001
Accuracy	0.811	0.909	0.905–0.913	0.820	0.810–0.831	<0.001
Sensitivity	0.922	0.937	0.931–0.943	0.809	0.794–0.825	<0.001
Specificity	0.700	0.889	0.882–0.895	0.833	0.814–0.852	<0.001
PPV (precision)	0.755	0.886	0.881–0.892	0.831	0.813–0.849	<0.001
NPV	0.899	0.939	0.932–0.945	0.814	0.796–0.832	<0.001
F1 score	0.830	0.908	0.904–0.912	0.817	0.805–0.828	<0.001

Values for the deep neural network and for logistic regression were obtained by repeated cross-validation and represent the grand means of the specific performance measures; p values comparing deep learning to regression are reported.

Discussion

Using data from 140 patients, we have demonstrated the feasibility of predicting GTR in transsphenoidal surgery for PAs using deep learning with excellent precision. Our deep learning model outperformed both the Knosp classification and logistic regression.

The decision to operate or not depends on various factors. In nonfunctioning adenomas, visual field loss, hypopituitarism, headaches, and growth of an incidentaloma are the usual surgical indications.[5,17] In secreting adenomas, biochemical cure through GTR is often the surgical goal, especially for patients with Cushing's syndrome or acromegaly.[18,23] Predictive analytics for GTR may be useful in risk stratification, as well as in surgical decision-making whenever there is no unequivocal indication for surgery.

The Knosp classification has existed since 1993 and has seen widespread use due to its relative simplicity and accuracy.[7,11,14] This is corroborated by our findings, which indicate that the Knosp classification alone is already a valuable predictive tool that provides adequate accuracy and sensitivity. The reported deep learning model outperformed the Knosp classification and logistic regression. The improvement in accuracy is explained not only by the technique of machine learning itself, but also by the greater number of variables that the deep learning model is able to integrate. In addition to other machine learning methods, deep learning is able to discover and abstract several levels of interactions between variables.

For example, the Knosp classification gains its predictive ability only from the degree of parasellar extension, whereas our deep learning model was also able to combine this information with data on, for example, suprasellar extension derived from tumor and sellar measurements in 3 axes, demographic factors, and overall volume.

Conventionally, classifications often struggle with predicting the risk of bad outcomes or complications particularly in "moderate" cases, which roughly corresponds to Knosp grades 2 and 3A in PAs. Advanced predictive analytics could be of particular use in these moderate cases. This is well illustrated in Fig. 2. Low-grade and high-grade adenomas were almost all correctly predicted by all 3 methods. However, deep learning outperformed both the dichotomized Knosp classification and logistic regression for grade 2 and 3A adenomas.

Logistic regression models are widely used to predict outcomes. The trained regression model demonstrated good predictive ability. In fact, logistic regression is a machine learning method itself, and most artificial neural networks are built on the same principles as logistic regression. However, logistic regression itself does not have the same aptitude as deep learning for reducing many complex relationships between features into multiple simpler problems.

The main advantage of conventional statistical models, like regression, over deep learning lies within interpretability.[2,19] Many machine learning models, including deep neural networks, represent black boxes where an input and output are known, but understanding their internal decision-making process is not feasible. Due to the output of odds ratios and p values, conventional statistical methods such as logistic regression are much easier to interpret. Deep learning thus enables higher predictive accuracy at the cost of reduced interpretability.

As with any statistical or machine learning model, overfitting is a possible problem when applying the trained model to external data. Overfitting occurs when a model adjusts too closely to the data that it is given for training, which produces very high accuracy on the training data

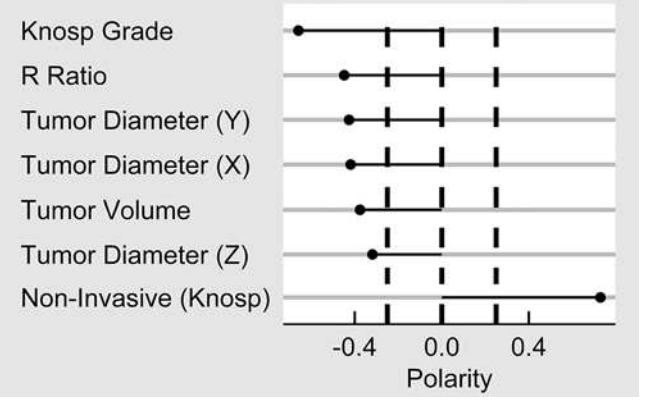

FIG. 1. To demonstrate which variables were most valuable to the deep neural network, a polarity correlation plot was constructed. Negative polarities indicate an inverse correlation between a certain feature and GTR. A greater deviation from zero implies higher feature importance.

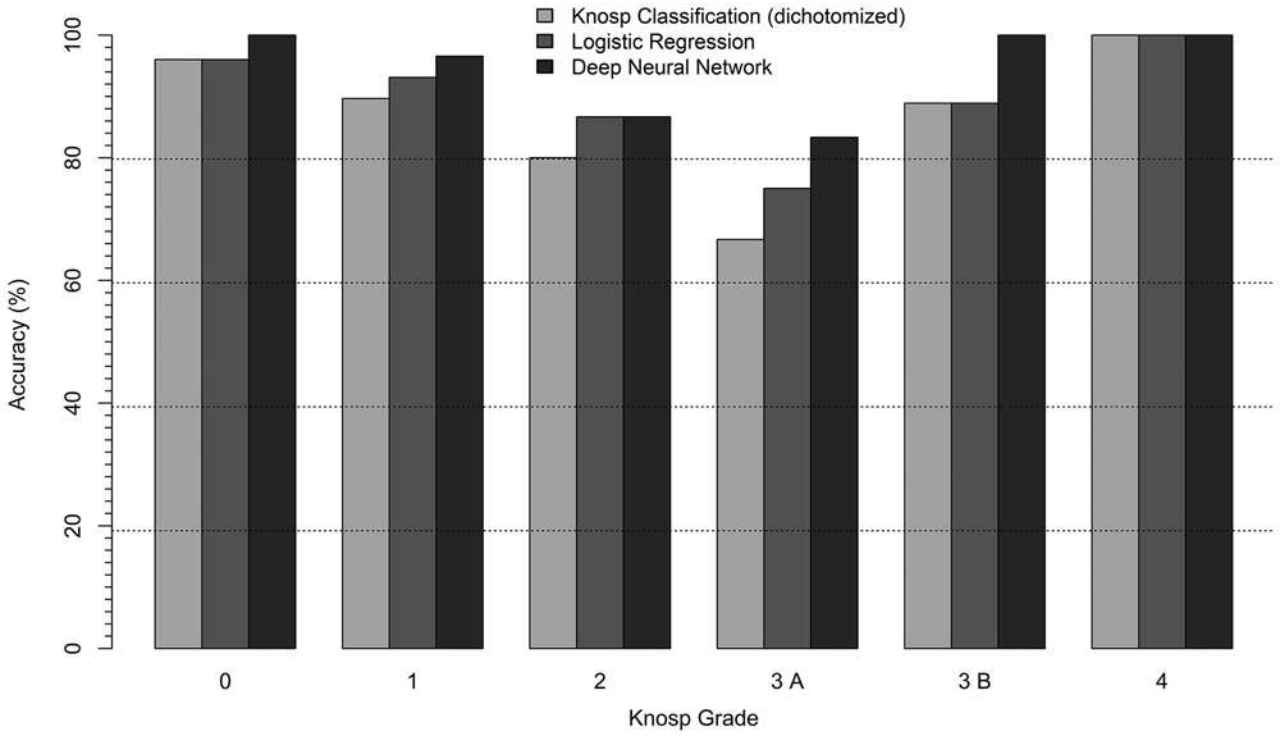

FIG. 2. Bar graph demonstrating the accuracy of the 3 prediction models for GTR per modified Knosp grade. Grades 2 and 3A are clearly more difficult to predict for all of the prediction tools. However, deep neural networks appear to outperform the other methods in these "moderate" cases. For this illustration, predictions were made on the entire set of 140 patients.

set, but subsequently results in poor performance on testing or external validation data. In other words, the model simply starts memorizing the training data instead of learning the interactions between features and endpoints. We effectively prevented overfitting using the validated dropout technique.[22] Furthermore, we assessed out-of-sample error using *k*-fold cross-validation without holdout, which allows us to say that our model should show similar performance on comparable external data. Often, 20% of the data are kept as a holdout set for final testing of the model. However, in this pilot study, we opted for conventional cross-validation without holdout due to our already small training data set, and because we wanted to statistically compare deep learning with logistic regression, which required repeated cross-validation. The trained logistic regression model, however, is probably overfitted since we did not employ any specific way to prevent overfitting here.

In contrast to other machine learning methods, deep learning has not previously been extensively used for neurosurgical outcome prediction. In a systematic review, Senders et al. found that machine learning models generally augment the decision-making capacity of clinicians, but that developing, validating, and deploying these techniques into daily clinical practice is often difficult.[19] Azimi et al. also reviewed the literature for applications of artificial neural networks in neurosurgery and found that they, among other things, are effective at diagnosis of low-back pain, brain tumors, and epilepsy; interpreting brain and spine imaging; and predicting outcomes in a range of diseases.[2] However, they reported that 49 of the 50 identified studies used "conventional" neural networks, indicating that deep learning has not arrived in neurosurgical research and practice as of yet. In other fields, deep learning often outperforms other machine learning methods by a large margin.[12] Our study is one of the very first to apply deep learning methods to neurosurgical outcome prediction and shows that the introduction of this technique has the potential to provide relevant real-world improvement in predictive medicine.

The primary limitation of our study is its retrospective nature, although all data were entered into a prospective registry. In regard to the deep learning model, the moderate sample size may represent a limitation. We applied powerful and validated data science and augmentation techniques to extract the maximum amount of information from our data. GTR was attempted even in cases deemed invasive whenever safely possible, but not in all cases. While this is certainly part of routine clinical practice, it may affect our findings. In addition, all data stem from a single center, which may limit the generalizability of our model.

In this pilot study, we demonstrated that deep learning can be used to predict GTR in a robust fashion and with improved performance in comparison with already precise gold standards. However, before our model can be deployed for use in daily clinical practice, it must be trained on a larger sample from a prospective multicenter study. Currently, our model requires 16 different input variables to attain its high performance. This would be cumbersome

for clinicians. A larger sample will allow further improving and validating performance, while scaling down the amount of required inputs to just a few easily measured variables, making it easy to use in daily clinical practice.

Conclusions

In this pilot study, we demonstrated the utility of applying deep learning to preoperatively predict the likelihood of GTR with excellent performance. An easy-to-use deep learning model would be a valuable addition to risk stratification and surgical decision-making.

References

1. Asher AL, Devin CJ, Archer KR, Chotai S, Parker SL, Bydon M, et al: An analysis from the Quality Outcomes Database, Part 2. Predictive model for return to work after elective surgery for lumbar degenerative disease. **J Neurosurg Spine 27:**370–381, 2017
2. Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A: Artificial neural networks in neurosurgery. **J Neurol Neurosurg Psychiatry 86:**251–256, 2015
3. Bouthillier A, van Loveren HR, Keller JT: Segments of the internal carotid artery: a new classification. **Neurosurgery 38:**425–433, 1996
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic minority over-sampling technique. **J Artif Intell Res 16:**321–357, 2002
5. Dallapiazza RF, Grober Y, Starke RM, Laws ER Jr, Jane JA Jr: Long-term results of endonasal endoscopic transsphenoidal resection of nonfunctioning pituitary macroadenomas. **Neurosurgery 76:**42–53, 2015
6. Dehdashti AR, Ganna A, Karabatsou K, Gentili F: Pure endoscopic endonasal approach for pituitary adenomas: early surgical results in 200 patients and comparison with previous microsurgical series. **Neurosurgery 62:**1006–1017, 2008
7. Dhandapani S, Singh H, Negm HM, Cohen S, Anand VK, Schwartz TH: Cavernous sinus invasion in pituitary adenomas: systematic review and pooled data meta-analysis of radiologic criteria and comparison of endoscopic and microscopic surgery. **World Neurosurg 96:**36–46, 2016
8. Elhadi AM, Hardesty DA, Zaidi HA, Kalani MYS, Nakaji P, White WL, et al: Evaluation of surgical freedom for microscopic and endoscopic transsphenoidal approaches to the sella. **Neurosurgery 11 (Suppl 2):**69–79, 2015
9. Hardy J, Vezina JL: Transsphenoidal neurosurgery of intracranial neoplasm. **Adv Neurol 15:**261–273, 1976
10. Kanter AS, Dumont AS, Asthagiri AR, Oskouian RJ, Jane JA Jr, Laws ER Jr: The transsphenoidal approach. A historical perspective. **Neurosurg Focus 18(4):**e6, 2005
11. Knosp E, Steiner E, Kitz K, Matula C: Pituitary adenomas with invasion of the cavernous sinus space: a magnetic resonance imaging classification compared with surgical findings. **Neurosurgery 33:**610–618, 1993
12. LeCun Y, Bengio Y, Hinton G: Deep learning. **Nature 521:**436–444, 2015
13. Meij BP, Lopes MBS, Ellegala DB, Alden TD, Laws ER Jr: The long-term significance of microscopic dural invasion in 354 patients with pituitary adenomas treated with transsphenoidal surgery. **J Neurosurg 96:**195–208, 2002
14. Micko ASG, Wöhrer A, Wolfsberger S, Knosp E: Invasion of the cavernous sinus space in pituitary adenomas: endoscopic verification and its correlation with an MRI-based classification. **J Neurosurg 122:**803–811, 2015
15. Mooney MA, Hardesty DA, Sheehy JP, Bird R, Chapple K, White WL, et al: Interrater and intrarater reliability of the Knosp scale for pituitary adenoma grading. **J Neurosurg 126:**1714–1719, 2017
16. Negm HM, Al-Mahfoudh R, Pai M, Singh H, Cohen S, Dhandapani S, et al: Reoperative endoscopic endonasal surgery for residual or recurrent pituitary adenomas. **J Neurosurg 127:**397–408, 2017
17. Przybylowski CJ, Dallapiazza RF, Williams BJ, Pomeraniec IJ, Xu Z, Payne SC, et al: Primary versus revision transsphenoidal resection for nonfunctioning pituitary macroadenomas: matched cohort study. **J Neurosurg 126:**889–896, 2017
18. Schwyzer L, Starke RM, Jane JA Jr, Oldfield EH: Percent reduction of growth hormone levels correlates closely with percent resected tumor volume in acromegaly. **J Neurosurg 122:**798–802, 2015
19. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al: Natural and artificial intelligence in neurosurgery: a systematic review. **Neurosurgery 83:**181–192, 2018
20. Serra C, Burkhardt JK, Esposito G, Bozinov O, Pangalu A, Valavanis A, et al: Pituitary surgery and volumetric assessment of extent of resection: a paradigm shift in the use of intraoperative magnetic resonance imaging. **Neurosurg Focus 40(3):**E17, 2016
21. Serra C, Maldaner N, Muscas G, Staartjes V, Pangalu A, Holzmann D, et al: The changing sella: internal carotid artery shift during transsphenoidal pituitary surgery. **Pituitary 20:**654–660, 2017
22. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R: Dropout: a simple way to prevent neural networks from overfitting. **J Mach Learn Res 15:**1929–1958, 2014
23. Sughrue ME, Chang EF, Gabriel RA, Aghi MK, Blevins LS: Excess mortality for patients with residual disease following resection of pituitary adenomas. **Pituitary 14:**276–283, 2011
24. Zaidi HA, De Los Reyes K, Barkhoudarian G, Litvack ZN, Bi WL, Rincon-Torroella J, et al: The utility of high-resolution intraoperative MRI in endoscopic transsphenoidal surgery for pituitary macroadenomas: early experience in the Advanced Multimodality Image Guided Operating suite. **Neurosurg Focus 40(3):**E18, 2016

Disclosures

Dr. Regli: consultant for BB Braun Medical.

Author Contributions

Conception and design: Serra, Staartjes, Regli. Acquisition of data: all authors. Analysis and interpretation of data: Serra, Staartjes, Regli. Drafting the article: Serra, Staartjes. Critically revising the article: Serra, Regli. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Serra. Statistical analysis: Staartjes. Administrative/technical/material support: Serra, Regli. Study supervision: Serra, Regli.

Supplemental Information

Online-Only Content

Supplemental material is available online.

Appendix 1. https://thejns.org/doi/suppl/10.3171/2018.8.FOCUS18243.

Correspondence

Carlo Serra: University Hospital Zürich, Zürich, Switzerland. c.serra@hotmail.it.